US007996390B2

(12) United States Patent
Freire et al.

(10) **Patent No.: US 7,996,390 B2**
(45) **Date of Patent: Aug. 9, 2011**

(54) **METHOD AND SYSTEM FOR CLUSTERING IDENTIFIED FORMS**

(75) Inventors: **Juliana Freire**, Salt Lake City, UT (US); **Luciano Barbosa**, Salt Lake City, UT (US)

(73) Assignee: **The University of Utah Research Foundation**, Salt Lake City, UT (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 385 days.

(21) Appl. No.: **12/032,331**

(22) Filed: **Feb. 15, 2008**

(65) **Prior Publication Data**

US 2009/0210406 A1 Aug. 20, 2009

(51) **Int. Cl.**
G06F 7/00 (2006.01)
G06F 17/30 (2006.01)

(52) **U.S. Cl.** **707/722**; 707/723; 707/728; 707/736; 707/737; 707/758

(58) **Field of Classification Search** 707/705, 707/720–735, 999.1, 999.3, 999.6, 999.7, 707/736, 737, 758
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,754,873	B1	6/2004	Law et al.	
7,080,073	B1	7/2006	Jiang et al.	
7,213,198	B1	5/2007	Harik	
2002/0169770	A1 *	11/2002	Kim et al.	707/5
2003/0220912	A1 *	11/2003	Fain et al.	707/3
2005/0097436	A1 *	5/2005	Kawatani	715/500
2005/0203924	A1 *	9/2005	Rosenberg	707/100
2006/0129446	A1	6/2006	Ruhl et al.	
2006/0200478	A1	9/2006	Pasztor	
2006/0230033	A1 *	10/2006	Halevy et al.	707/3
2007/0100812	A1	5/2007	Simske et al.	
2007/0100862	A1	5/2007	Reddy et al.	
2007/0112898	A1 *	5/2007	Evans et al.	707/205
2008/0154942	A1 *	6/2008	Tsai et al.	707/102

OTHER PUBLICATIONS

Huang et al., "Multi-type Features Based Web Document Clustering", Springer Berlin / Heidelberg, vol. 3306/2004, pp. 253-265, 2004. Download: http://www.springerlink.com/content/te7qn81416wqy7g6/.*

Barbosa et al., "Searching for Hidden-Web Database", Eighth International Workshop on the Web and Database, Jun. 16-17, 2005, pp. 1-6. Download: http://webdb2005.uhasselt.be/papers/1-1.pdf.*

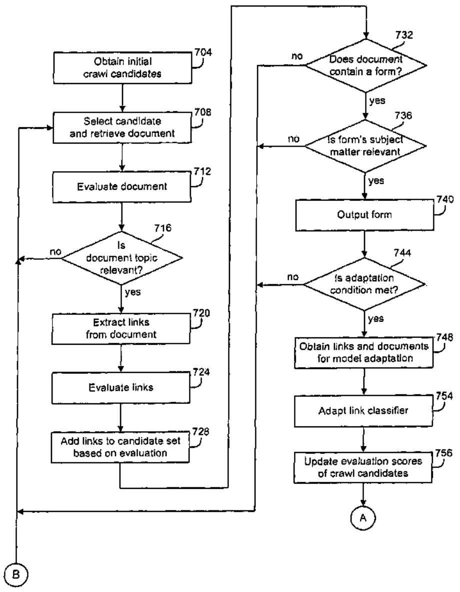
(Continued)

Primary Examiner — John E Breene
Assistant Examiner — Hares Jami
(74) *Attorney, Agent, or Firm* — Bell & Manning, LLC

(57) **ABSTRACT**

A method is provided for organizing a plurality of documents that include forms. An initial set of clusters is defined for the plurality of documents. The initial set of clusters is reclustered based on similarity values calculated in multiple feature spaces. For example, a first feature space may be associated with a content of a document while a second feature space may be associated with a content of a form associated with the document. Each cluster has an associated centroid vector in each feature space that is used to represent the cluster. The similarity between the document and each cluster is calculated in both feature spaces. Each document is assigned to the cluster whose centroid is most similar. The cluster centroids may be recalculated and the process repeated until the cluster assignments become stable.

32 Claims, 10 Drawing Sheets


Obtain initial crawl candidates
704
Select candidate and retrieve document
708
Evaluate document
712
Is document topic relevant?
716
no
yes
Extract links from document
720
Evaluate links
724
Add links to candidate set based on evaluation
728
732
Does document contain a form?
no
yes
736
Is form's subject matter relevant
no
yes
Output form
740
744
Is adaptation condition met?
no
yes
Obtain links and documents for model adaptation
748
Adapt link classifier
754
Update evaluation scores of crawl candidates
756
A
B